Tune-An-Ellipse: CLIP Has Potential to Find What You Want

Jinheng Xie[1] Songhe Deng[2] Bing Li[3] Haozhe Liu[3] Yawen Huang[4†] Yefeng Zheng[4]
Jürgen Schmidhuber[3] Bernard Ghanem[3] Linlin Shen[2] Mike Zheng Shou[1†]
[1] Show Lab, National University of Singapore [2] Shenzhen University
[3] AI Initiative, King Abdullah University of Science and Technology
[4] Jarvis Research Center, Tencent YouTu Lab
{sierkinhane,mike.zheng.shou}@gmail.com
https://github.com/showlab/Tune-An-Ellipse

Abstract

Visual prompting of large vision language models such as CLIP exhibits intriguing zero-shot capabilities. A manually drawn red circle, commonly used for highlighting, can guide CLIP's attention to the surrounding region, to identify specific objects within an image. Without precise object proposals, however, it is insufficient for localization. Our novel, simple yet effective approach, i.e., Differentiable Visual Prompting, enables CLIP to zero-shot localize: given an image and a text prompt describing an object, we first pick a rendered ellipse from uniformly distributed anchor ellipses on the image grid via visual prompting, then use three loss functions to tune the ellipse coefficients to encapsulate the target region gradually. This yields promising experimental results for referring expression comprehension without precisely specified object proposals. In addition, we systematically present the limitations of visual prompting inherent in CLIP and discuss potential solutions.

1. Introduction

Large Language Models (LLMs) [2, 17, 19] and Vision-Language Models (VLMs) [18, 20] emerged in recent years are featured with some fascinating properties such as *prompting* [1, 13, 31, 40]. Notably, in the absence of predefined supervision signals during training, LLMs exhibit a robust zero-shot capability across diverse tasks such as translation and question answering, simply through prompting. For instance, when prompted by "Translate them into English", LLMs can accordingly translate the user's input. VLMs such as CLIP [20] have demonstrated similar ability, in which the prompt of "a photo of {}" can enable zero-shot image classification and boost performance. In light of this, many works [5, 9, 39] have been dedicated to finding better text prompts for VLMs, with strategies like CoOp [39] to equip the prompt with learnable tokens.

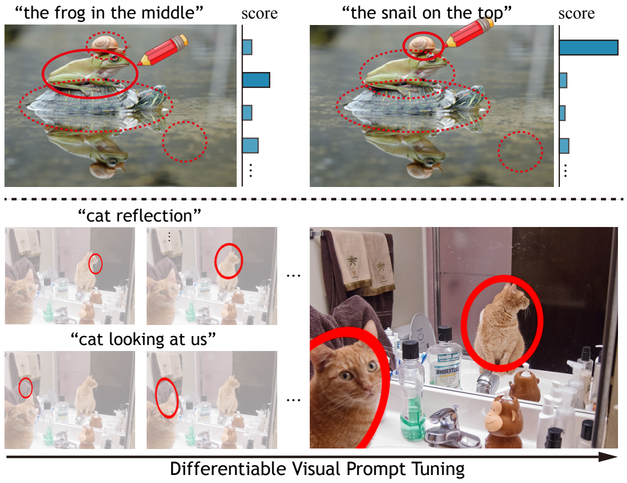


Figure 1. **Top**: An illustration of visual prompt for CLIP [20]. As observed in [24], the red circle can guide CLIP's attention towards the surrounding region. For example, given a text prompt "the snail on the top", the CLIP score reaches its maximum when the circle is drawn around the snail. **Bottom**: Our Tune-An-Ellipse. Given a text prompt, we begin by generating an initial ellipse by visual prompting and propose three loss functions to iteratively tune the ellipse coefficients to encapsulate the target region.

Unlike previous efforts purely focusing on the textual prompts, Aleksandar *et al.* [24] observed that a visually prompted image—a red circle simply drawn around a region—allows CLIP to pinpoint any specific instances from an image that may contain multiple objects. For an intuitive explanation, people commonly use a red circle to highlight and emphasize important elements, and this visual prompt is then implicitly learned by CLIP through contrastive learning on web-scale data. As the example shown in Fig. 1, given a text prompt "the frog in the middle" and various visually prompted images, CLIP can correspondingly infer a distinct matching score for the proper prompted

[†] Corresponding Author

one. Beyond object recognition, Aleksandar *et al.* [24] also showed that scoring a set of *pre-extracted object proposals* by visual prompting could empower CLIP to localize objects described in a referring expression [16, 34]. However, this approach hinges on the precise object proposals, typically obtained from external models like Faster R-CNN [21] with limited pre-defined concepts. While more advanced open-world detectors can be integrated, the range of their conceptual knowledge still lags behind that of CLIP. Thus, enabling CLIP itself to localize objects described in referring expressions is valuable but remains challenging.

Many works such as CAM [38] and Grad-CAM [22] have devoted effort to explain what regions the neural networks would attend to [3, 12, 22, 29, 38]. These methods employ learned weights or gradients to generate an activation map that vividly shows the model's attention. Fig. 2 (c) illustrates an example of Grad-CAM, in which we show an activation map of CLIP in response to the expression "the yellow fish at the bottom". One can observe that while Grad-CAM can correctly identify the relevant regions, it also attends to considerable inaccurate content. This observation suggests that Grad-CAM alone is insufficient to accurately localize specific objects due to its tendency to produce erroneous attention.

This paper proposes to enable CLIP for zero-shot referring expression comprehension (REC) based on visual prompting and Grad-CAM, without requiring precise object proposals. The core idea is to iteratively tune an initial ellipse to encapsulate the target region accurately via a neural network. Specifically, given an image and a textual prompt, we initially generate an ellipse with coefficients via the proposed differentiable visual prompting and activation map of Grad-CAM from a set of uniformly distributed anchor ellipses (as shown in Figs. 2 (a) and (b)). Subsequently, the coefficients follow the equation of a rotated ellipse to form an ellipse curve, which can be embedded in the image as a visual prompt (Figs. 2 (e) and (b)). This manner makes the visual prompting process differentiable such that gradient descent can be employed to adjust the initial coefficients. With tailored learning objectives, a neural network is adopted to predict proper transformations, moving the ellipse curve to encapsulate the target region. We illustrate the process and some examples at the bottom of Fig. 1. As observed, given an image and a text prompt "cat looking at us", an initial ellipse curve near the cat is generated, then will be iteratively tuned by the neural network to completely and compactly encapsulate the cat looking at us.

In addition to presenting our proposed approach, we systematically analyze the limitations of visual prompting within CLIP, which currently serves as a bottleneck for more practical applications. We also engage in a discussion regarding potential avenues for enhancing CLIP's visual prompting capabilities.

We summarize the main contributions as follows:

- We propose a novel, simple yet effective approach, *i.e.,* differentiable visual prompting, to enabling CLIP itself to localize objects described in referring expressions without precisely specified object proposals.
- We systematically present the limitations of visual prompting inherent in CLIP and discuss potential avenues for further improvement.

2. Related Works

Prompting in Large Scale Models. Benefiting from scaling parameters and corpus, GPT series [2, 17, 19] exhibit remarkable zero-shot ability in tasks like translation and chatting with a handful of prompts or samples. For instance, by prompted "Translate the following English text to French", these models can suggest an accurate French translation for the subsequent input. Indeed, apart from the models' vast parameter space, this feature may mainly stem from the inclusion of prompts that encapsulate human intent within the training data. Likewise, beyond language corpus, recent studies [24] have shown that large vision-language models like CLIP can also learn and understand human intent in web-scale visual data, such as images with red circles highlighting something important. Aleksandar *et al.* [24] pointed out that such prompts effectively direct CLIP's attention to a specific area within an image, facilitating zero-shot recognition for a selected instance instead of a rough overview.

Visual Prompt Tuning. Nurtured by large-scale multimodal data from the web, VLMs such as CLIP have the capacity to comprehend massive concepts both in textual and visual forms. However, it is worth noting that CLIP may not yet exhibit expertise in certain downstream tasks. To mitigate this issue, many visual prompt tuning approaches on text input [5, 9, 39], vision input [7, 28, 33, 37], or both text and vision inputs [23, 36] have been proposed to efficiently and effectively adapt CLIP to specific tasks. For instance, Zhou *et al.* [39] proposed to learn a set of learnable tokens as context words to the textual input by fine-tuning on a few labeled images in specific tasks or domains.

Referring Expression Comprehension (REC). Given a textual description, REC aims to find the image region most relevant to the expression. Commonly, most works [6, 14, 15, 30, 32, 35] focus on identifying the corresponding region by scoring a set of pre-extracted object proposals (usually generated by Faster R-CNN [21]). For instance, Yu *et al.* [35] proposed a modular attention network to decompose expression into three modular components, in which two kinds of attention are employed to score the object proposals. Recently, researchers proposed to tackle referring expression comprehension in an unsupervised / zero-shot manner [8, 24, 26, 33]. ReCLIP [26] first isolates object proposals via cropping and blurring and em-

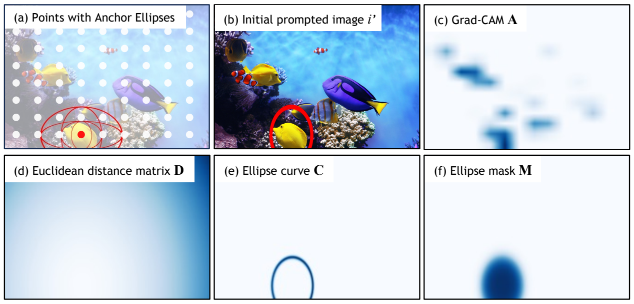

(a) Points with Anchor Ellipses
(b) Initial prompted image i'
(c) Grad-CAM $\mathbf{A}$
(d) Euclidean distance matrix $\mathbf{D}$
(e) Ellipse curve $\mathbf{C}$
(f) Ellipse mask $\mathbf{M}$

Figure 2. An illustration of the differentiable visual prompting process. The text prompt is "the yellow fish at the bottom".

ploys CLIP to score each proposal. In addition, a spatial relation resolver is leveraged to mitigate the weakness of CLIP on spatial reasoning. CPT [33] paints the object proposals with various colors and gets the most relevant colored proposal via a pre-trained captioning model. More recently, Aleksandar *et al.* [24] showed the capability of CLIP to recognize visual markers such as a red circle, and then proposed to prompt the image via such visual markers and get the matched proposals via CLIP scoring.

3. Methodology

The goal is to tune a visual prompt, *i.e.,* a red ellipse ○, to get a maximum overlap with the region most relevant to the text prompt. This work builds upon CLIP [20], a model designed to align image and text information in the same feature space. Following the previous study [24], we denote the input image as $i \in \mathbb{R}^{H\times W\times 3}$ and text as $t \in \Sigma^*$, where Σ is the alphabet. CLIP is hereby formulated as a function $s(\cdot)$ to predict a score $s(i,t) \in \mathbb{R}^1$, which measures the semantic similarity between i and t. In addition, we denote a fixed set of background text prompts [11] such as "a clean origami land" and "a clean origami wall" as $\{t_1^-, t_2^-, \cdots\}$. Furthermore, the activation map obtained by applying Grad-CAM on image-text pair (i,t) is denoted as $\mathbf{A} \in \mathbb{R}^{H\times W}$.

In the following, we first introduce how to make the process of visual prompting differentiable (§ 3.1) and then elaborate on the proposed learning objectives to tune the initial ellipse toward the target region (§ 3.2).

3.1. Differentiable Visual Prompting

The visual prompt ○ can be represented as a rotated ellipse $\phi(x,y)$ parameterized by the ellipse center (c_x, c_y), major axis a, minor axis b, and rotated angle θ:

$$\begin{aligned}\phi(x,y) = &\frac{((x-c_x)\cos\theta + (y-c_y)\sin\theta)^2}{a^2} \\ &+ \frac{((x-c_x)\sin\theta - (y-c_y)\cos\theta)^2}{b^2} - 1,\end{aligned} \tag{1}$$

where $x, y : \Omega \subset \mathbb{R}^2$. As shown in Fig. 2 (d), given a mesh grid to $\phi(\cdot)$, a matrix of Euclidean distance $\mathbf{D} \in \mathbb{R}^{H\times W}$ can be derived, where the distance values on the ellipse curve are all zeros. In this way, we can transform $\mathbf{D}$ into a matrix $\mathbf{C}$ with an approximate ellipse curve on it by applying an un-normalized Gaussian distribution $f(\cdot)$:

$$f(\phi(x,y)) = \exp\left(-\frac{(\phi(x,y)-\mu)^2}{2\sigma^2}\right), \tag{2}$$

where μ and σ^2 represent the mean and variance. By setting $\mu = 0$ and a proper σ, a rotated ellipse curve can be approximated and used to visually prompt the image i to i' (as shown in Figs. 2 (e) and (b)).

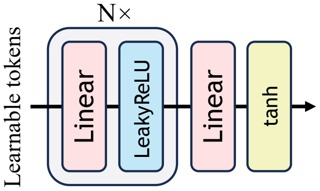


Figure 3. The architecture of MLP model.

In addition, we also transform $\mathbf{D}$ to a mask of approximately binarized rotated ellipse $\mathbf{M}$ using the following modulated inverse tangent function $g(\cdot)$ with ϵ:

$$g(\phi(x,y)) = \frac{1}{2}\left(1 + \frac{2}{\pi} arctan\left(\frac{\phi(x,y)}{\epsilon}\right)\right), \tag{3}$$

where ϵ controls the fuzziness. Such a transformation is illustrated from Fig. 2 (d) to Fig. 2 (f). Note that $g(\cdot)$ can be any function, *e.g.,* Sigmoid function, capable of approximating the Heaviside step function.

3.2. Ellipse Coefficients Tuning

Initialization. Here, we introduce how to obtain a relatively good initialization of the ellipse coefficients to the target region. Motivated by anchor-based object detection [21] and SAM [10], we uniformly sample N points on the image and assign each point with M various anchor ellipses in different sizes. We showcase some examples in Fig. 2 (a). For an image, we can totally obtain NM proposals, resulting in a set of visually prompted images $\{i'_1, \cdots, i'_{NM}\}$. Then, we measure the similarity between each prompted image and the text input to get a set of scores $\{s(i'_1, t), \cdots, s(i'_{NM}, t)\}$ and calculate a set of average activation $\{\frac{\sum(\mathbf{A}_1\cdot\mathbf{M}_1)}{\sum\mathbf{M}_1}, \cdots, \frac{\sum(\mathbf{A}_{NM}\cdot\mathbf{M}_{NM})}{\sum\mathbf{M}_{NM}}\}$ with Grad-CAM. Subsequently, we initialize the ellipse coefficients by selecting the proposal with the highest average activation among the Top-K matching scores. We notate the initial visually prompted image as i^* with the ellipse coefficients of $(c_x^*, c_y^*, a^*, b^*, \theta^*)$, transformed binarized mask as $\mathbf{M}^*$ and the corresponding Grad-CAM as $\mathbf{A}^*$. Note that $\mathbf{A}^*$ is fixed over the optimization process.

MLP Model. Given a set of initial coefficients, we aim to tune them to approach the region most relevant to the text prompt. A small MLP model is employed to predict a set of

Algorithm 1 Ellipse Coefficients Tuning

```
Preparation:
A set of initial ellipse coefficients $(c_x^*, c_y^*, a^*, b^*, \theta^*)$.
Prompted image $i^*$, text and background prompts $t, t_j^-$.
Main loop:
for $step \in \{1, \dots, max\_step\}$ do
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{inf}} + \mathcal{L}_{\text{sqz}}$
    Calculate gradients $\nabla\mathcal{L}$ to update the MLP model.
    Predict $(t_x, t_y, t_a, t_b, t_\theta)$ using the MLP model.
    Use $(t_x, t_y, t_a, t_b, t_\theta)$ to update $(c_x^*, c_y^*, a^*, b^*, \theta^*)$.
    Get the updated prompted image $i^*$.
end for
```

transformations $(t_x, t_y, t_a, t_b, t_\theta)$ to shift the initial location to $(c_x+t_x, c_y+t_y, a+t_a, b+t_b, \theta+t_\theta)$. The overall model architecture is presented in Fig. 3. It consists of stacked linear layers with LeakyReLU$(\cdot)$ activation, ending with a tanh$(\cdot)$. Note that a set of learnable tokens is employed as the input of the MLP model.

Optimization. To tune the current ellipse coefficients, we logistically design three loss functions. The first one is the matching loss $\mathcal{L}_{\text{sim}}$ based on the similarity scores:

$$\mathcal{L}_{\text{sim}} = -\log \frac{\exp(s(i^*, t))}{\exp(s(i^*, t)) + \sum_j \exp(s(i^*, t_j^-))}, \tag{4}$$

where t_j^- is the j-th background prompt. $\mathcal{L}_{\text{sim}}$ maximizes the similarity between the visually prompted image i^* and caption t, guiding the MLP model to predict transformations that align the ellipse curve with the target region. However, in principle, the concepts are not independent, and CLIP can, at times, derive relevant information from the background that aligns with the input caption. This implies that there exist local optima within the image space that can distract the attention of the network. To this end, inspired by mathematical morphology [25], we propose to dynamically adjust the size of the ellipse to prevent the model from getting trapped in local optimal solutions (background). This yields two additional learning objectives: inflation loss $\mathcal{L}_{\text{inf}}$ and squeezing loss $\mathcal{L}_{\text{sqz}}$. Specifically, $\mathcal{L}_{\text{inf}}$ is formulated as:

$$\mathcal{L}_{\text{inf}} = -\log\left(\frac{\sum(\mathbf{A}^* \cdot \mathbf{M}^*)}{HW}\right). \tag{5}$$

$\mathcal{L}_{\text{inf}}$ aims to inflate the red ellipse curve to include more activation. However, as Grad-CAM often attends to those irrelevant regions, the involvement of $\mathcal{L}_{\text{inf}}$ may over-inflate the ellipse curve. To mitigate the problem, we introduce the squeezing loss $\mathcal{L}_{\text{sqz}}$:

$$\mathcal{L}_{\text{sqz}} = -\log\left(\frac{\sum(\mathbf{A}^* \cdot \mathbf{M}^*)}{\sum \mathbf{M}^*}\right). \tag{6}$$

$\mathcal{L}_{\text{sqz}}$ maximizes the average activations inside the rotated ellipse, equivalent to encouraging the ellipse to cover the target region compactly.

A combination of these learning objectives would result in a well-suited ellipse curve covering the target region:

$$\mathcal{L} = \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{inf}} + \mathcal{L}_{\text{sqz}}, \tag{7}$$

where the weights of three loss terms are set as 1. We present the coefficients tuning process in Algorithm 1.

4. Experiments

4.1. Experimental Setup

Datasets and Evaluation Metrics. Following the protocol [24, 26], we evaluate the proposed method using the referring expression comprehension (REC) task, which aims to find the image region that is most relevant to the given text input. The REC task is commonly evaluated on RefCOCO [34], RefCOCO+ [34], and RefCOCOg [16], in which each image is annotated with multiple expressions and each expression refers to a unique object with bounding box information. In particular, expressions in RefCOCO and RefCOCOg include relation-based words such as left/bigger/closer, and only appearance-based descriptions are involved in RefCOCO+. For RefCOCO and RefCOCO+, it is split into *testA* and *testB* for people and non-people evaluation. We evaluate performance using the percentage of accurate predictions, considering a box as correctly predicted when its intersection-over-union with the ground-truth box exceeds 0.5.

Implementation Details. Akin to prior studies [24, 26], we employ an ensemble of two CLIP vision encoders (ViT-B/16 and ViT-L/14) to compute the loss, and an ensemble of ViT-B/16 and ViT-L/14@336px to get the Grad-CAM in our experiments. More experimental results on various CLIP vision encoders are provided in supplementary materials. Given an image i, text prompt t, and background prompts $\{t_1^-, t_2^-, \cdots\}$, we can accordingly get a list of scores $\{s(i,t), s(i,t_1^-), s(i,t_1^-), \cdots\}$ and then stack them together to obtain a vector. We further apply the Softmax function to normalize them and get the activation map of Grad-CAM following the open-source implementation[1]. In the process of differentiable visual prompting, σ in Eq. 2 is set as 0.05. The number of nodes in each MLP layer is 64 (input), 128, 128, and 5 (output), respectively. The learnable tokens are randomly initialized. To ensure the MLP model predicts a more stable transformation vector, we multiply the input of tanh$(\cdot)$ by 0.5. We adopt the Adam optimizer and set the learning rate as 0.001 with a cosine annealing scheduler. The default tuning steps are set as 200.

[1] https://github.com/jacobgil/pytorch-grad-cam

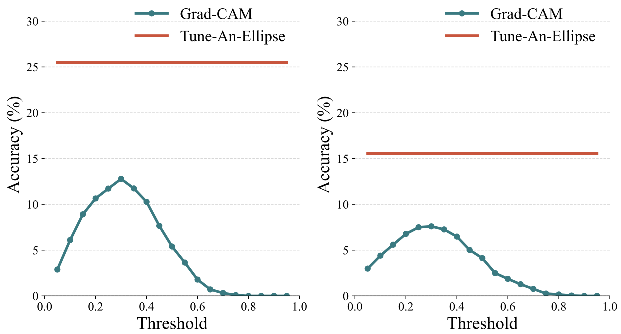
Methods	ZS	Proposal	RefCOCO			RefCOCO+			RefCOCOg	
			val	*testA*	*testB*	*val*	*testA*	*testB*	*val*	*test*
DTWREG [27]	×	F-RCNN	39.2	41.1	37.7	39.2	40.1	38.1	–	–
Pseudo-Q [8]	×	F-RCNN	56.0	58.3	54.1	38.9	45.1	32.1	46.3	47.4
CPT [33]	✓	F-RCNN	32.2	36.1	30.3	31.9	35.2	28.8	36.7	36.5
ReCLIP [26]	✓	F-RCNN	45.8	46.1	47.1	47.9	50.1	45.1	59.3	59.0
Red Circle* [24]	✓	F-RCNN	48.6	56.2	41.7	54.7	61.5	46.0	59.3	58.9
Ours*	✓	F-RCNN	49.8	58.0	40.9	55.1	62.8	45.7	58.4	58.7
ReCLIP [26]	✓	P-Ellipses	7.35	6.59	7.69	7.90	6.88	8.99	12.1	11.7
Red Circle [24]	✓	P-Ellipses	8.34	6.56	10.2	8.94	6.85	11.3	16.0	15.1
Ours	✓	P-Ellipses	**26.3**	**32.3**	**19.8**	**27.8**	**32.7**	**22.1**	**29.9**	**29.0**

Table 1. **Comparison with state-of-the-art on REC.** We present Top-1 accuracy (%). ZS refers to the zero-shot setting. Two types of proposals, *i.e.,* Faster R-CNN (F-RCNN) and points with anchor ellipses (P-Ellipses) (Fig. 2 (a)), are considered in the comparison. F-RCNN means that object-bounding boxes detected by Faster R-CNN are used as object proposals. P-Ellipses indicate that we just uniformly sample points on the image and use anchor ellipses of different sizes as the proposals (as introduced in Sec. 3.2). Red Circle* indicates that results are based on our reimplementation. Ours* means that we score the object proposals the same as Red Circle but use the proposed differentiable visual prompting.

4.2. Quantitative Results

We present a comparison with state-of-the-art REC approaches in Table 1. Zero-shot REC methods generally initialize a set of candidate object proposals detected by Faster R-CNN, score each proposal, and then select the proposal with the highest score as the final localization. In this way, CPT [33], ReCLIP [26] and Red Circle [24] achieve competitive performance to those supervised methods such as DTWREG [27] and Pseudo-Q [8]. A potential limitation of this approach may lie in their performance drop when encountering expressions with un-predefined concepts like ”snail” that the Faster R-CNN detector fails to localize. As shown in Table 1, regarding uniformly distributed anchor ellipses as object proposals, these methods significantly fail to localize the target region, reaching low accuracy across RefCOCO, RefCOCO+, and RefCOCOg datasets. For example, ReCLIP only obtains 8.3%, 6.5%, and 10.2% accuracy on RefCOCO *val*, *testA*, and *testB*, respectively. In contrast, our Tune-An-Ellipse can relatively mitigate the problem by tuning the proposals to encapsulate the target region, which can achieve more promising results. As observed, our approach can get better accuracy of 26.3%, 32.3%, and 19.8% on RefCOCO *val*, *testA*, and *testB*, respectively.

As Grad-CAM is involved in the proposed loss terms, we present a comparison with it on RefCOCO *testA* and *testB* in Fig. 4. For a fair comparison, we evaluate accuracy at the pixel level, considering a mask as correctly predicted when its intersection over-union with the ground-truth mask exceeds 0.5. By applying a threshold, binary masks can be derived from Grad-CAM activation maps. Notably, Grad-CAM exhibits sensitivity to threshold variations, leading to significant performance fluctuations between thresholds of 0 and 1. In contrast, the proposed Tune-An-Ellipse does not require the thresholding process and always achieves better performance than Grad-CAM.



(a) RefCOCO *testA*

(a) RefCOCO *testB*

Figure 4. Comparison with Grad-CAM on the referring expression comprehension (REC) task. We utilize an ensemble of ViT-B/16 and ViT-L/14 for our method and Grad-CAM. To ensure a fair comparison, we derive binary masks from Grad-CAM by applying a threshold and calculating the mask accuracy using the ground-truth mask. Similarly, we transform the predicted ellipse into a mask to compute the mask accuracy for our method.

4.3. Visualization Results

We showcase the visualization of referring expression comprehension results for Tune-An-Ellipse in Fig. 5. The samples are drawn from RefCOCO *val*, *testA*, and *testB*. The ground truth is represented by green bounding boxes, while the predicted results from Tune-An-Ellipse are depicted using red ellipses. One can observe that Tune-An-Ellipse can accurately find the target region described in the referring expressions, even in some complicated scenes. For instance, as observed in the top-right of Fig. 5, though the lady and guy stand in the background of the two salient players, given referring expressions “lady middle pink” and “guy in back left red shirt”, the proposed Tune-An-Ellipse can also precisely and completely find the corresponding regions. These cases highlight the significant potential of vanilla CLIP, coupled with our tuning approach, in accurately interpreting expressions and finding the corresponding target regions, without the need for precise object proposals extracted by pre-trained detectors.

Visualizations of Grad-CAM and the localization results of Tune-An-Ellipse are presented in Fig. 6. Grad-CAM often focuses on irrelevant or incorrect regions, posing challenges in localizing objects described in referring expressions alone. For example, in the top-left corner of Fig. 6, when provided with the referring expression “adult male center right”, Grad-CAM erroneously focuses more on the head of the child, accompanied by background activation. Moreover, Grad-CAM frequently tends to concentrate on sub-words within an expression, as illustrated by the emphasis on the “hat” in the top-right of Fig. 6. Additionally, it often highlights discriminative parts of the target object,

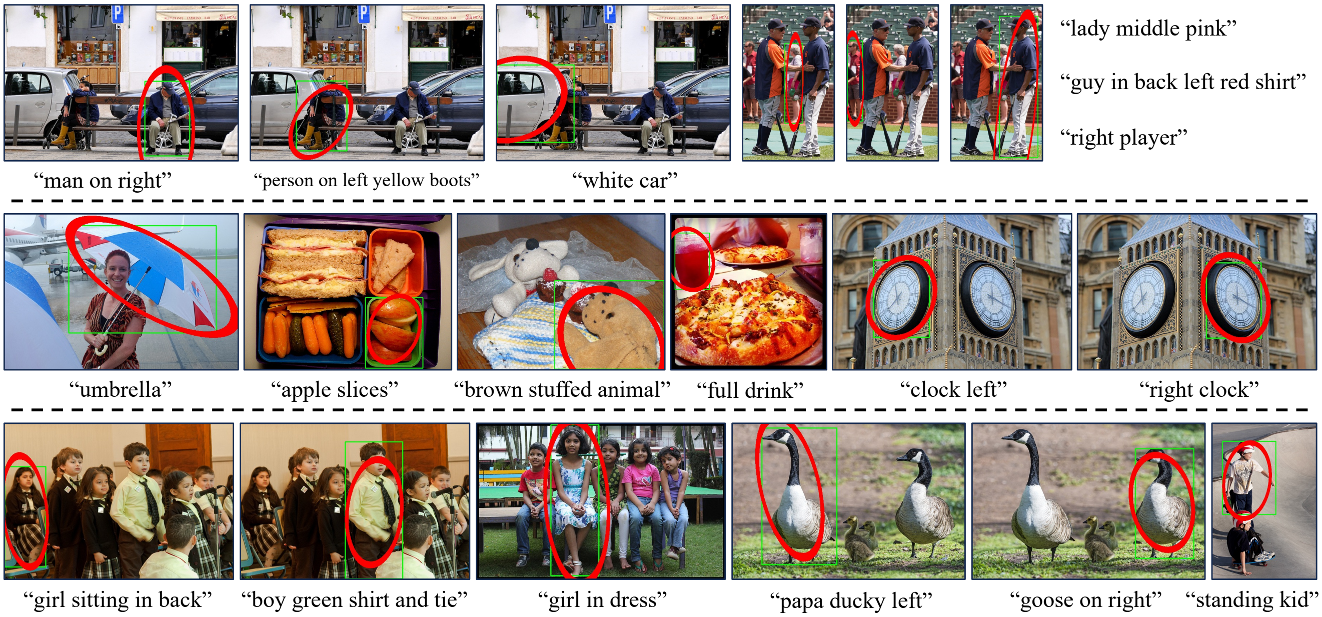


Figure 5. Localization results of Tune-An-Ellipse on RefCOCO *val*, *testA*, and *testB*. The bounding boxes in green are the ground truth and the red ellipses are the results predicted by the proposed Tune-An-Ellipse. The texts such as "papa ducky left" are the referring expressions.

as seen with the focus on the head of the person in the middle right of Fig. 6. In contrast, Tune-An-Ellipse effectively avoids incorrect attention by visually prompting anchor ellipses and subsequently tuning the initial ellipse to accurately and completely encapsulate the target regions. For example, in the bottom-left of Fig. 6, when presented with the expression "front orange", Tune-An-Ellipse accurately locates the region of the orange in the front, despite the presence of erroneous activation around the right orange.

4.4. Ablation Studies

As introduced in Sec. 3.2, $\mathcal{L}_{\text{inf}}$ and $\mathcal{L}_{\text{seq}}$ complement each other to prevent the tuned ellipse from over-inflating or underestimating the target regions. It may seem that these two loss terms alone are sufficient for the accurate localization of target objects. In practice, the $\mathcal{L}_{\text{sim}}$ is necessary and we illustrate its importance in Fig. 6, which presents visualization of Grad-CAM, localization results of Tune-An-Ellipse w/o $\mathcal{L}_{\text{sim}}$ on RefCOCO *testA* and *testB*. It is evident that Grad-CAM often focuses on incomplete or incorrect regions that are irrelevant to the referring expressions. For example, in the top-right of Fig. 6, when given the expression "man with hat", Grad-CAM exclusively attends to the region of the hat, disregarding the entire body of the man. With only $\mathcal{L}_{\text{inf}}$ and $\mathcal{L}_{\text{seq}}$, the resulting red ellipse only covers the hat. Upon incorporating $\mathcal{L}_{\text{seq}}$, the final predicted red ellipse accurately and compactly encompasses the man with a hat. Furthermore, Grad-CAM may erroneously focus on inaccurate regions, leading to significant localization inaccuracies. For instance, in the bottom-right of Fig. 6, when presented with the expression "laying elephant", most of the Grad-CAM activation concentrates on the standing elephant rather than the laying one. The absence of $\mathcal{L}_{\text{sim}}$ results in notably inaccurate localization, as evidenced by similar instances in the middle row. These examples underscore the necessity and effectiveness of $\mathcal{L}_{\text{sim}}$. More numerical results on ablation studies of the three loss terms $\mathcal{L}_{\text{inf}}$, $\mathcal{L}_{\text{seq}}$, and $\mathcal{L}_{\text{sim}}$ are presented in Table 2. It is evident that when all three loss terms are involved, Tune-An-Ellipse achieves the best performance. Moreover, an ensemble of Grad-CAM with two CLIP vision encoders (simply adding two activation maps) results in improved performance.

Ablation studies on various initialization manners and tuning steps are presented in Table 3. Without any tuning, the initial ellipses can only obtain around 4% accuracy. With the proposed tuning method, it can achieve significant improvement with an accuracy of 32.3%. In addition, the table shows that selecting the anchor ellipse with the highest average activation among the Top10 matching scores, *i.e.,* Top10 (S) & Top1 (A), would lead to better performance.

We perform the sensitivity analysis of the number of Top-K, uniform points, and anchor ellipses as object proposals introduced in Sec. 3.2 and the results are presented in Table 4. It is evident that, with a fixed number of points (9*9) and anchor ellipses (6), opting for the Top10 anchor ellipses with the highest scores and subsequently selecting the Top-1 ellipse with the largest average activation yields a better performance on RefCOCO. When the number of K and anchor ellipses is fixed at 10 and 6, respectively, employing 9*9 uniformly sampled points across the image results in a better performance on RefCOCO. Besides, when the number of K and points is fixed at 10 and 9*9, respec-

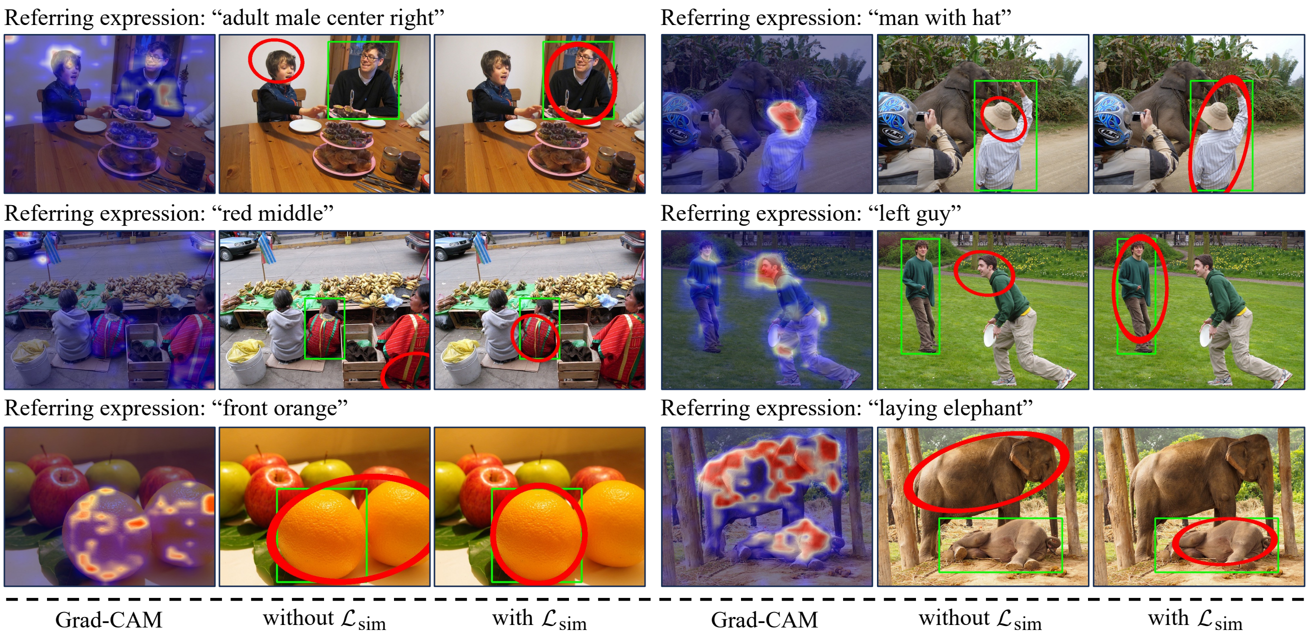

Referring expression: “adult male center right”
Referring expression: “man with hat”
Referring expression: “red middle”
Referring expression: “left guy”
Referring expression: “front orange”
Referring expression: “laying elephant”
Grad-CAM
without $\mathcal{L}_{\text{sim}}$
with $\mathcal{L}_{\text{sim}}$
Grad-CAM
without $\mathcal{L}_{\text{sim}}$
with $\mathcal{L}_{\text{sim}}$

Figure 6. Visualization of Grad-CAM and localization results of Tune-An-Ellipse w/o $\mathcal{L}_{\text{sim}}$ on RefCOCO *testA* and *testB*. Ground truth bounding boxes are indicated in green, and the results predicted by Tune-An-Ellipse are depicted by red ellipses.

$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{inf}}$	$\mathcal{L}_{\text{seq}}$	Accuracy (%)
✓			4.63 / 5.30
	✓		5.78 / 5.75
		✓	0.14 / 0.19
✓	✓		6.01 / 6.28
✓		✓	0.25 / 0.48
	✓	✓	31.2 / 27.2
✓	✓	✓	**32.3 / 29.0**

Table 2. Impact of three losses on REC Top-1 accuracy based on ensemble Grad-CAM from CLIP ViT-B/16 and ViT-L/14@336 or based on single Grad-CAM. The results are from RefCOCO *testA*.

tively, assigning 6 anchor ellipses for each point yields a better performance on RefCOCO.

Initialization	Tuning Steps	Accuracy (%)
Top-1 (S)	0	3.94
Top-1 (A)	0	1.26
Top-1 (S)	200	29.1
Top-1 (A)	200	31.6
Top-10 (S) & Top-1 (A)	0	3.13
Top-10 (A) & Top-1 (S)	0	3.96
Top-10 (A) & Top-1 (S)	200	31.8
Top-10 (S) & Top-1 (A)	200	**32.3**

Table 3. Ablation studies on initialization manners and tuning steps. Top-1 (S) indicates the anchor ellipse with Top-1's highest matching score and Top-1 (A) indicates the anchor ellipse with Top-1's largest average activation. Top-10 (S) & Top-1 (A) indicates the anchor ellipse with the highest average activation among the Top-10 matching scores.

#K	#Points	#Anchor Ellipses	Accuracy (%)
5	9 * 9	6	31.6
15	9 * 9	6	31.8
10	6 * 6	6	31.5
10	12 * 12	6	31.8
10	9 * 9	3	31.3
10	9 * 9	6	**32.3**
10	9 * 9	9	31.8

Table 4. Sensitivity analysis *w.r.t.* the number of Top-K, grid points, and anchor ellipses introduced in Sec. 3.2.

5. Limitations

Deficiency of CLIP's Visual Prompting. Note that the visual prompting ability of CLIP sometimes still falls short of perfection, especially in situations without precise object proposals. This potentially lowers the upper bound of our method. Here, we systematically list the challenges below.

(a) **High Response to Background**. Without precise object proposals, many anchor ellipses are falling in the background, such as sky and grass. Unfortunately, a red ellipse drawn on the background can sometimes lead to a high response of CLIP to the referring expressions. We showcase examples in Fig. 7 (a). Given “the big one”, the visual prompt ○ on the background, such as sky and grass, would get a higher matching score than that of well around the ground truth. These cases underscore that the current CLIP's visual prompting ability might encounter challenges in effectively eliminating background interference.

(b) **High Response to Partial Inclusion**. REC strives for precise and complete localization of referring objects. Unfortunately, CLIP does not possess such an ability perfectly as illustrated in Fig. 7 (b). Given “right bear”, the red ellipses that completely encompass the target objects receive matching scores of 0.47 and 0.30, respectively, even

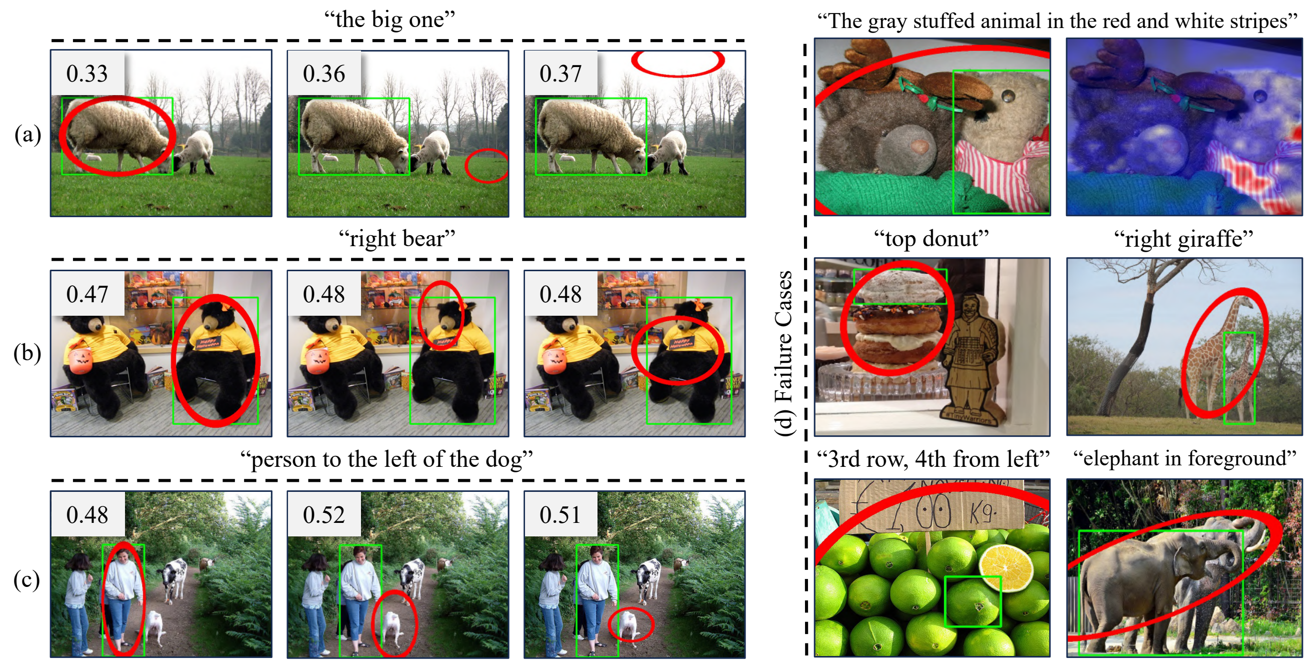


Figure 7. Illustration of scenarios where CLIP's visual prompting falls short. We categorize three main scenarios: (a) high response to the background, (b) high response to inaccurate inclusion, and (c) high response to sub-words of expressions. The bounding boxes in green are the ground truth. **Number on the left corner of the image is the matching score** $s(i^*, t)$ between visually prompted image i^* and text prompt t. The scores are the summation of two CLIPs, *i.e.,* ViT-B/16 and ViT-L/14.

lower than that of partially including the target object.

(c) **High Response to Sub-words**. In REC, the task involves a comprehensive understanding of expressions followed by the accurate localization of corresponding regions. Unfortunately, CLIP sometimes makes mistakes in this situation as shown in Fig. 7 (c). When presented with "person to the left of the dog", the goal is to localize the person. However, CLIP gives the visual prompt of the dog on the right a higher score than that of the person.

Potential Solutions. We attribute the above challenges to the limited representation of both positive and negative examples of such visual prompts during CLIP's training. In addition to incorporating more visually prompted images to train vision-language models, introducing visual prompts on the background and part of the target object as negative samples for contrastive learning would potentially enhance CLIP's visual prompting ability to mitigate the problems.

6. Discussions

As mentioned above, CLIP's visual prompting ability may not be flawless in certain cases, necessitating the use of additional loss terms such as $\mathcal{L}_{\text{inf}}$ and $\mathcal{L}_{\text{sqz}}$ based on Grad-CAM for compensation. However, these terms can potentially lead to the problem of over-tuning, since Grad-CAM is also not reliable as depicted in Fig. 7 (d). We argue that if CLIP's visual prompting ability is perfect, only the proposed differentiable visual prompting with $\mathcal{L}_{\text{sim}}$ would be sufficient to accurately localize the target within the given context, which will be extremely simple and effective.

7. Conclusion

Building upon the emerging visual prompting capability of CLIP, this work proposed an approach, *i.e.,* differentiable visual prompting, to enabling CLIP to automatically localize the image region most relevant to referring expressions, eliminating the requirement of precise object proposals from detectors such as Faster R-CNN. Our approach involved the generation of an initial ellipse from uniformly distributed anchor ellipses through visual prompting. Subsequently, three loss functions were employed to iteratively refine the ellipse coefficients, gradually encapsulating the target region. Experimental results demonstrated that our method can achieve promising results in zero-shot referring expression comprehension. In addition, we also systematically outlined the challenges of visual prompting within CLIP and engaged in a discussion regarding potential avenues for improvement.

Acknowledgment

Mike Zheng Shou and Xie Jinheng were supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008. Bing Li, Haozhe Liu, and Bernard Ghanem were supported by the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, pages 1877–1901, 2020. 1

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, pages 1877–1901, 2020. 1, 2

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, pages 839–847, 2018. 2

[4] Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. gscorecam: What objects is clip looking at? In *ACCV*, pages 1959–1975, 2022. 1

[5] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *CVPR*, pages 2808–2817, 2023. 1, 2

[6] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016. 2

[7] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 2

[8] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *CVPR*, pages 15513–15523, 2022. 2, 5

[9] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124, 2022. 1, 2

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023. 3

[11] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*, pages 15305–15314, 2023. 3

[12] Haozhe Liu, Mingchen Zhuge, Bing Li, Yuhui Wang, Francesco Faccio, Bernard Ghanem, and Jürgen Schmidhuber. Learning to identify critical states for reinforcement learning from videos. In *ICCV*, pages 1955–1965, 2023. 2

[13] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35, 2023. 1

[14] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. Learning cross-modal context graph for visual grounding. In *AAAI*, pages 11645–11652, 2020. 2

[15] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *CVPR*, pages 7102–7111, 2017. 2

[16] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 2, 4

[17] OpenAI. Chatgpt. 2023. 1, 2

[18] OpenAI. Gpt-4 technical report. 2023. 1

[19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, page 9, 2019. 1, 2

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 3

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, pages 91–99, 2015. 2, 3

[22] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 2, 1

[23] Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E Gonzalez, Kurt Keutzer, and Trevor Darrell. Multitask vision-language prompt tuning. *arXiv preprint arXiv:2211.11720*, 2022. 2

[24] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, pages 11987–11997, 2023. 1, 2, 3, 4, 5

[25] Pierre Soille et al. *Morphological image analysis: principles and applications*. 1999. 4

[26] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, pages 5198–5215, 2022. 2, 4, 5

[27] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Yannis Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *TPAMI*, 43:4189–4195, 2021. 5

[28] Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In *CVPR*, pages 7725–7735, 2023. 2

[29] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPRW*, pages 24–25, 2020. 2, 1

[30] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*, page 1960–1968, 2019. 2

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, pages 24824–24837, 2022. 1

[32] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *ICCV*, pages 4643–4652, 2019. 2

[33] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 2, 3, 5

[34] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 2, 4

[35] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018. 2

[36] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 2

[37] Sheng Zhang, Salman H. Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *CVPR*, pages 3479–3488, 2023. 2

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 1, 2

[40] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*, 2023. 1